Computational Models of Relevance Propagation in Web Directories*

Eduardo Xamena[†§] Nélida Beatriz Brignole[†§] Ana G. Maguitman[‡**]

[†] LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica
[‡] LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Av. Alem 1253, (8000) Bahía Blanca, Argentina
phone: 54-291-4595135 fax: 54-291-4595136

[§]Planta Piloto de Ingeniería Química (UNS-CONICET)
Cno la Carrindanga km 7, (8000) Bahía Blanca, Argentina

e-mail: examena@plapiqui.edu.ar {nbb,agm}@cs.uns.edu.ar

Abstract. Web Directories consist of large collections of links to websites, arranged by topic in different categories. The structure of Web Directories is typically not flat, since there are hierarchical and non-hierarchical relationships among topics. As a consequence, websites classified under certain topic may be relevant to other topics. While some of these relevance relations are explicit, most of them must be discovered by an analysis of the structure of these directories. This paper proposes a family of models of relevance propagation in Web Directories. An efficient computational framework for inferring implicit relevance relations is described. The framework presented here combines the use of matrices to represent relevance relations and the application of boolean operations on these matrices to infer implicit relations. Nine different models were computed for a portion of the Open Directory Project graph consisting of more than half a million nodes and approximately 1.5 million edges of different types. The models were compared by performing both a quantitative and qualitative analysis on them. It is found that some general difficulties rule out the possibility of defining flawless models of relevance propagation that only take into account structural features of Web Directories.

1 Introduction

A Web Directory is a directory of webpages classified by topic into categories. Examples of Web Directories are Yahoo! Directory[1], Open Directory Project (ODP)[2], and their derivatives, such as Google Directory[3]. While regular Web search is the most common way adopted by users to find information on a specific topic, Web Directories are particularly useful to navigate through related topics, or when the user is not sure how

* This research work is partially supported by CONICET (PIP 11220090100863) and Universidad Nacional del Sur (PGI 24/N026 and PGI 24/ZN13).

** To whom correspondence should be addressed.

[1] http://dir.yahoo.com.

[2] http://dmoz.org.

[3] http://www.google.com/dirhp.

to narrow her or his search from a broad category. Web Directories can help understand how topics within a specific area are related and may suggest terms that are useful in conducting a search. Besides being organized by topic, webpages classified in these directories have the advantages of having annotations (such as a description) and having been evaluated by an editor. ODP, for instance, has 20,000 volunteer editors reviewing websites and classifying them by topic.

Although Web Directories were originally conceived as a means to organize webpages to facilitate its navigation by humans, the content and structure of these directories are increasingly being used to serve other purposes. For instance, Google's regular Web search results are enhanced by information from Google Directory. ODP has been used to train and test automatic classifiers [4,9], as the starting point to collect thematic material by topical crawlers [7,15], as a framework to understand the structure of content-based communities on the Web [6], to implement Information Retrieval evaluation platforms [3,12], to understand the evolution of communities in P2P search [1], and to evaluate the emergent semantics of social tagging [14], among other applications. Many of these applications rely on identifying relevance or semantic similarity relationships between webpages classified in ODP.

Relevance is a powerful concept employed in various subdisciplines within Computer Science, especially in Artificial Intelligence and Information Science. An initial analysis of the problem of defining the relevance between documents classified in a Web Directory indicates that it essentially involves the problem of identifying non-obvious relationships from the directory structure. Identifying these relationships in Web Directories is a challenging problem. The structure of Web Directories is typically not flat since topics can be classified according to some taxonomic schema. Topic taxonomies contain parent-child relationships between topics and their subtopics. However, relationship schemes other than parent-child hierarchies are also common. For example, the ODP ontology is more complex than a simple tree. Some topics have multiple criteria to classify subtopics. The "Business" category, for instance, is subdivided by types of organizations (cooperatives, small businesses, major companies, etc.) as well as by areas (automotive, health care, telecom, etc.). Furthermore, ODP has various types of cross-reference links between categories, so that a node may have multiple parent nodes, and even cycles are present. The combination of different kinds of links gives rise to intricate relations among topics. While some of these relations are explicitly given by the existing links most of them remain implicit. Currently, ODP contains more than one million categories, making the problem of automatically deriving implicit relations between topics computationally very hard.

The goal of this paper is twofold: (1) to present a family of computational models to efficiently derive implicit relationships among topics from the structure of these directories, and (2) to analyze the limitations of these models and discuss ways to overcome them.

2 Representing the Structure of a Web Directory Graph

A Web Directory Graph is a directed graph of nodes representing topics. Each node contains objects representing documents (webpages). A Web Directory Graph has a

hierarchical (tree) component made by "is-a" links, and non-hierarchical components made by cross links of different types.

For example, the ODP ontology is a directed graph $G = (V, E)$ where:

- V is a set of nodes, representing topics containing documents;
- E is a set of edges between nodes in V, partitioned into three subsets T, S and R, such that:
  - T corresponds to the hierarchical component of the ontology,
  - S corresponds to the non-hierarchical component made of "symbolic" cross links,
  - R corresponds to the non-hierarchical component made of "related" cross links.

Figure 1 shows a simple example of a Web Directory Graph extracted from ODP. In this graph, the set V contains topic nodes such as REFERENCE, EDUCATION, SCHOOL_SAFETY, LABS_AND_EXPERIMENTS, etc. The subset T corresponding to the hierarchical component of the Web Directory Graph contains edges such as (TOP,REFERENCE), (REFERENCE,EDUCATION), (EDUCATION,SCHOOL_SAFETY), etc. In this example there is a "symbolic" edge: (SCIENCE_FAIRS,SCIENCE) and two "related" edges: (LABS_AND_EXPERIMENTS,SCIENCE_FAIRS) and (SCIENCE,PUZZLES).

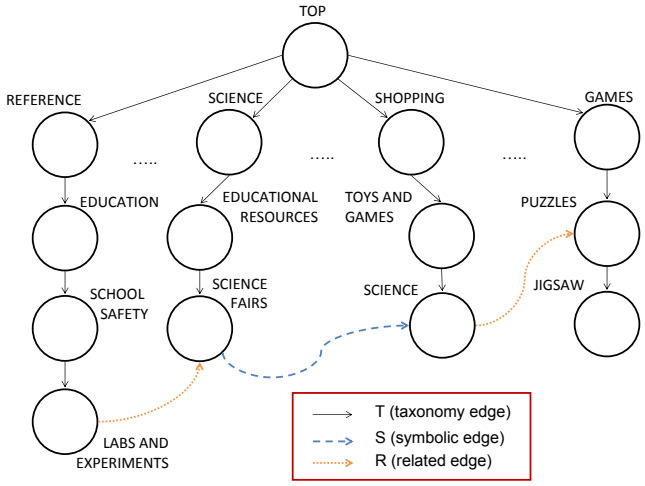

Fig. 1. Illustration of a Web Directory Graph extracted from ODP.

As a starting point, we say that topic t_j is relevant to topic t_i if there is an edge of some type from topic t_i to topic t_j. In the Web Directory Graph from figure 1, we can say that the topic SCHOOL_SAFETY is relevant to the topic EDUCATION, or that the topic SCIENCE_FAIRS is relevant to the topic LABS_AND_EXPERIMENTS, among other examples.

However, to derive implicit (indirect) topic relevance relations, transitive relations between edges should also be considered. An analysis of some examples leads us to conclude that while relevance relations are consistently preserved through hierarchical

links, it is necessary to impose certain constraints on how the non-hierarchical links can participate in the transitive relations. Allowing an arbitrary number of cross links is infeasible because it would relate each topic to almost every other topic. Take for example the portion of ODP shown in 1. In this example there is a path involving three edges between topics REFERENCE/EDUCATION/SCHOOL_SAFETY/LABS_AND_EXPERIMENTS and GAMES/PUZZLES but the relevance of the second topic to the first one is questionable. On the other hand, there are other indirect paths that preserve relevance, as is the case for the path of length three between SHOPPING/TOYS_AND_GAMES and GAMES/PUZZLES/JIGSAWS.

The question addressed here is: Can we automatically derive non-obvious relevance relations among topics? Our goal is to impose certain constraints on how cross links can participate in each path in such a way that we capture the non-hierarchical components of a Web Directory Graph while preserving meaning.

In order to build our computational models of relevance propagation we start by numbering the topics in V as t_1, t_2, ..., t_n, and by representing the Web Directory Graph structure by means of adjacency matrices. A matrix $\mathbf{T}$ is used to represent the hierarchical structure of an ontology. Matrix $\mathbf{T}$ codifies edges in T and is defined as $\mathbf{T}_{ij} = 1$ if $(t_i, t_j) \in T$ and $\mathbf{T}_{ij} = 0$ otherwise. The non-hierarchical components corresponding to the "symbolic" and "related" edges of the ODP graph are represented by matrices $\mathbf{S}$ and $\mathbf{R}$, respectively. Matrix $\mathbf{S}$ is defined so that $\mathbf{S}_{ij} = 1$ if $(t_i, t_j) \in S$ and $\mathbf{S}_{ij} = 0$ otherwise. The matrix $\mathbf{R}$ is defined analogously, as $\mathbf{R}_{ij} = 1$ if $(t_i, t_j) \in R$ and $\mathbf{R}_{ij} = 0$ otherwise.

3 Models of Relevance Propagation

Having codified the different components of the ODP graph as matrices $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{R}$, we proceed to address the question of how these matrices can be used to capture the notion of relevance. In the following, we present different models of relevance propagation and analyze some of their properties.

3.1 Explicit Relevance Relations

Consider the logical $\vee$ operation on matrices, defined as $[\mathbf{A} \vee \mathbf{B}]_{ij} = \mathbf{A}_{ij} \vee \mathbf{B}_{ij}$, and let $\mathbf{M_1}$ be computed as follows:

$$\mathbf{M_1} = \mathbf{T} \vee \mathbf{S} \vee \mathbf{R} \vee \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. Matrix $\mathbf{M_1}$ is the adjacency matrix of graph G augmented with 1s on the diagonal. While matrix $\mathbf{M_1}$ accounts for all the explicit relevance relations existing in ODP it fails to capture many indirect relevance relations that result from applying transitive closures or combining relations of different types.

3.2 Transitive Closure on the Hierarchical Component

In order to compose relations, we will use the boolean product of matrices defined as follows:

$$[\mathbf{A} \otimes \mathbf{B}]_{ij} = \bigvee_k (\mathbf{A}_{ik} \wedge \mathbf{B}_{kj}).$$

Let $\mathbf{T}^{(0)} = \mathbf{I}$, and let $\mathbf{T}^{(r+1)} = \mathbf{T} \otimes \mathbf{T}^{(r)}$.

Matrix $\mathbf{T}^{(r)}$ codifies all the paths of length r between topics. We define the closure of $\mathbf{T}$, denoted $\mathbf{T}^*$, as follows:

$$\mathbf{T}^* = \bigvee_{r=0}^{\infty} \mathbf{T}^{(r)}$$

Matrix $\mathbf{T}^*$ codifies all the paths (of any length) existing between pairs of topics following "is-a" links. Since there is a finite number of topics, matrix $\mathbf{T}^*$ can be computed in a finite number of steps. In this matrix, $\mathbf{T}^*_{ij} = 1$ if t_j belongs to the the topic subtree rooted at t_i, and $\mathbf{T}^*_{ij} = 0$ otherwise.

Since we have observed that relevance relations are consistently preserved through the "is-a" links it is reasonable to compute the transitive closure $\mathbf{T}^*$ and augment it with the matrices representing the "symbolic" and "related" links. This gives rise to our second model of relevance:

$$\mathbf{M_2} = \mathbf{T}^* \vee \mathbf{S} \vee \mathbf{R}.$$

In this new model, topic t_j is relevant to topic t_i if (1) there is a path from topic t_i to topic t_j involving "is-a" links only, or (2) there is a "symbolic" or "related" link from topic t_i to topic t_j. Model $\mathbf{M_2}$ is a conservative model in the sense that it propagates relevance through the hierarchical component of the ODP graph only, while the participation of cross-links is restricted to explicit (direct) relevance relations.

A question that arises next is whether cross links can be included in indirect paths while preserving meaning. We have observed earlier (figure 1) that relevance is often lost if an arbitrary number of cross links are added to a path. Therefore, for the relevance propagation models to be plausible certain constraints should be imposed.

Below we formulate a family of plausible models of relevance propagation, which result from extending the previous models.

3.3 Propagating Cross-Links throughout the Taxonomy

A simple way to incorporate cross links into the model is by propagating them upwards or downwards through the taxonomy. If we want to propagate relevance relations induced by cross links towards the root, we obtain the following model of relevance:

$$\mathbf{M_3} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}).$$

Alternatively, if we propagate relevance relations induced by cross links towards the leaves of the taxonomy we obtain the following model:

$$\mathbf{M_4} = (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

Finally, we can propagate relevance relations induced by cross links throughout *all* the taxonomy, but allowing a single cross link in each path. This results in the following model:

$$\mathbf{M_5} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

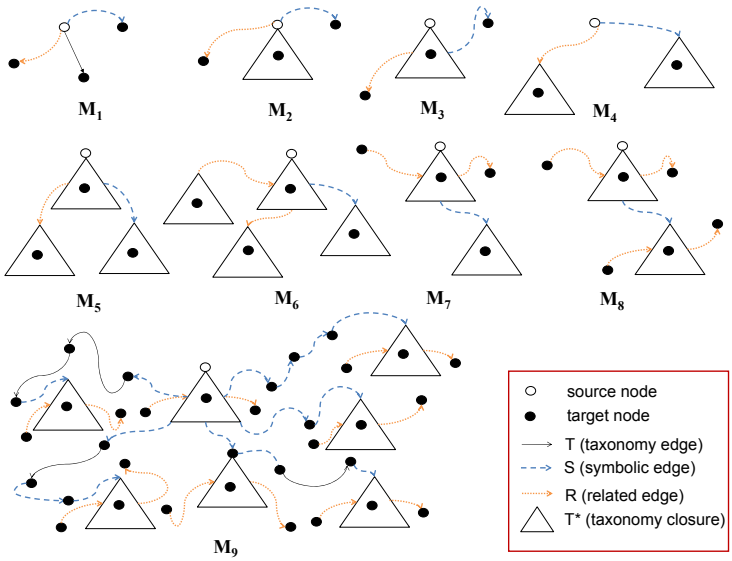

Fig. 2. Possible paths from source node to target nodes in different models of relevance propagation.

In previous work, model $\mathbf{M_5}$ of relevance propagation has been applied in the computation of semantic similarity measures with good results [13].

Another question that arises is whether relevance relations should be symmetric. The hierarchical component of the ODP graph (i.e., "is-a" links) codifies relevance relations from a child topic to its parent topic that in most of the cases are non-symmetric. In the meantime, since duplication of URLs is disallowed, "symbolic" links are a way to represent multiple memberships, for example the fact that the pages in topic SHOPPING/TOYS_AND_GAMES/SCIENCE also belong to topic SCIENCE/EDUCATIONAL_RESOURCES/SCIENCE_FAIRS. Therefore, "symbolic" links also codify parent-child relationships which, as is the case with "is-a" links, are generally non-symmetric. On the other hand, "related" links appear to codify symmetric relevance relations. Consequently, a new model of relevance can be formulated by making the "related" links bidirectional. This is achieved by extending the set of cross-link matrices with $\mathbf{R^T}$, i.e., the transpose of $\mathbf{R}$, resulting in the following model of relevance propagation:

$$\mathbf{M_6} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

Alternative models can be obtained by imposing additional constraints or by relaxing some. In general, "related" links appear to be weaker than the other types of links. We can reflect this in a new model that results from disallowing the downward propagation of "related" links:

$$\mathbf{M_7} = (\mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^*) \vee (\mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I})).$$

A generalization of $\mathbf{M_7}$ is $\mathbf{M_8}$, where both "symbolic" and "related" links are allowed to simultaneously participate in the same path:

$$\mathbf{M_8} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}).$$

There is a plethora of ways in which these models can be constrained or amplified. For example, we could allow up to n "symbolic" links as is shown in the following generalization of $\mathbf{M_8}$:

$$\mathbf{M_9} = \mathbf{T}^* \otimes (\mathbf{T} \vee \mathbf{S} \vee \mathbf{I})^n \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}).$$

Figure 2 shows possible relevance paths from a source to a target node according to the different models. Various models have been considered, but the ones discussed above capture the most interesting or salient aspects of the notion of relevance propagation analyzed here.

4 Analyzing the Models

4.1 Quantitative Analysis

The proposed models were computed for the ODP ontology. The portion of the ODP graph we have used for our analysis consists of 571,148 topic nodes (only the WORLD and REGIONAL categories were discarded). The following table shows the size of the components of the graph used in our analysis.

Component	Size
V	571,148 nodes
T	571,147 edges
S	545,805 edges
R	380,264 edges

In order to quantitatively compare the different models, we looked at the number of relevance relations between pairs of topics induced by each model. This comparison is shown in the following table.

Model	Number of relations
$\mathbf{M_1} = \mathbf{T} \vee \mathbf{S} \vee \mathbf{R} \vee \mathbf{I}$	2,068,364
$\mathbf{M_2} = \mathbf{T}^* \vee \mathbf{S} \vee \mathbf{R}$.	5,502,581
$\mathbf{M_3} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I})$	7,072,930
$\mathbf{M_4} = (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*$	71,443,444
$\mathbf{M_5} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*$	170,573,370
$\mathbf{M_6} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}) \otimes \mathbf{T}^*$	174,534,253
$\mathbf{M_7} = (\mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^*) \vee (\mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}))$	14,177,359
$\mathbf{M_8} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I})$	16,915,322
$\mathbf{M_9} = \mathbf{T}^* \otimes (\mathbf{T} \vee \mathbf{S} \vee \mathbf{I})^n \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I})$ with $n = 4$	37,609,462

The above comparison table reveals a wide variation in the number of relevance relations induced by each model. In addition, we computed the number of differences among the models, and observed that for some pairs of models, such as $\mathbf{M_6}$ and $\mathbf{M_9}$, the number of differences is as large as 177,799,003.

4.2 Qualitative Analysis

Having observed that the models produced quantitatively different characterizations of the notion of relevance, we proceeded to perform an analysis of the quality of the relations induced by each.

An important theoretical observation is that the set of models form a partial order under the relation "$\mathbf{M_m} \leq \mathbf{M_n}$ if and only if $[\mathbf{M_m}]_{ij} = 1$ implies $[\mathbf{M_n}]_{ij} = 1$ for all i, j". The resulting partial order is depicted in figure 3 and can be easily shown to hold by analyzing the definition of each model as well as the definitions of the $\vee$ and $\otimes$ operators.[4]

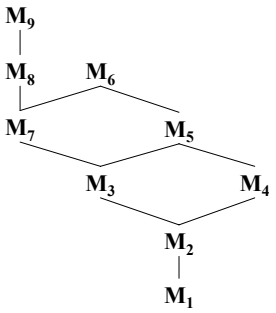

Fig. 3. Partial order on the set of models.

In order to dig deeper into the qualitative aspects of each model, we implemented a visualization tool. This tool was used in combination with the computed matrices to identify pairs of topics in which different models disagreed regarding the existence or absence of a relevance relation between them. Once conflicting topics were identified in the models, the visualization tool allowed us to visualize these topics and the set of webpages associated with them. This helped us to address the problem of which models produce the most accurate characterization of the notion of relevance.

Relevance is a highly subjective concept [5,2]. After an initial pilot experiment we observed low levels of agreement in relevance judgements between the human evaluators. To further complicate the task of evaluating the different models, we noticed that even for the same judge a relevance relation that existed at a certain point of time, may disappear later, or vice versa. Despite these discrepancies, for a good number of pairs of topics there was a clear agreement concerning the existence or absence of an implicit relevance relation. For example, in figure 1 the existence of an implicit relevance relation between the topic GAMES/PUZZLES/JIGZAW and the topic SHOPPING/TOYS_AND_GAMES is unquestionable, yet only models $\mathbf{M_5}$ and $\mathbf{M_6}$ capture this relation. On the other hand, there is not a clear relevance relation between the topics ARTS/ART_HISTORY/MOVEMENTS/IMPRESSIONISM and SOCIETY/ORGANIZATIONS/STUDENTS in figure 4 despite the facts that the less conservative models ($\mathbf{M_5}$, $\mathbf{M_6}$,$\mathbf{M_7}$,$\mathbf{M_8}$ and $\mathbf{M_9}$), would indicate the existence of such a relation.

[4] Furthermore, this is consistent with the models computed using the ODP graph.

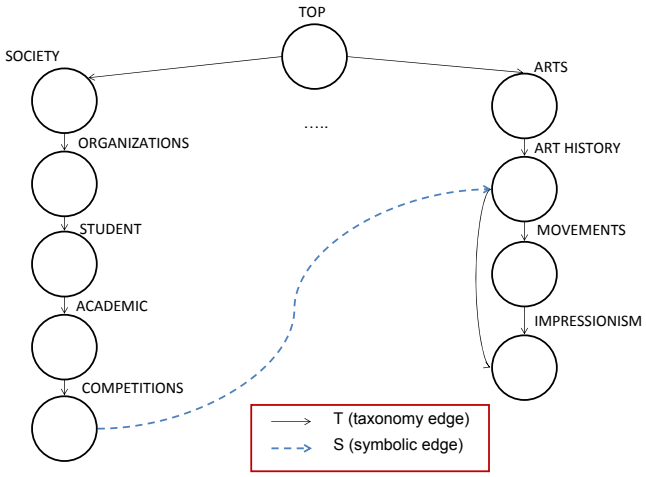

Fig. 4. A questionable relation in ODP.

Instances similar to the one illustrated in figure 4 are pervasive in ODP. This highlights the fact the less conservative schemes of relevance propagation are not robust because a few unreliable cross links make significant global changes to the relevance propagation models. On the other hand, the most conservative schemes are incomplete, and hence unable to derive many useful relevance relations induced by the less conservative ones.

5 Discussion

The above analysis leads us to conclude that while some models are better predictors than others of the existence or absence of relevance relations, none of them is flawless. This points to the fact that despite being a key concept in Artificial Intelligence and Information Science, relevance is a fuzzy and subtle notion, difficult, if not impossible, to formalize using structural aspects only.

There are a number of ways in which the proposed models of relevance propagation can be improved. For instance, the less conservative models could be combined with mechanisms that prevent them from deriving relevance relations between two topics unless an analysis of the topics' content suggests a connection between them. This analysis could be based on the text describing the topics, which is available in ODP. Another source of content are the features of the websites associated with the topics, such as the text, the outgoing links, the incoming link or a combination of all.

Another possible improvement is the extension of the proposed models to fuzzy models of relevance propagation. Different types of edges have different roles, and one way to distinguish these roles is to assign them weights. Then, the weight $w_{ij} \in [0, 1]$ for an edge between topic t_i and t_j can be interpreted as an explicit measure of the degree of membership of t_j in the family of topics rooted at t_i. In order to propagate relevance, the boolean product of matrices $\otimes$ will need to be replaced by some fuzzy

operator. For example, we could use the MaxProduct fuzzy composition operator [10] defined on matrices as follows:

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = \max_k (\mathbf{A}_{ik} \cdot \mathbf{B}_{kj}).$$

The element $\mathbf{M}_{ij}$ resulting from propagating relevance in the new fuzzy models will be interpreted as a fuzzy relevance relation of topic t_j to topic t_i. For ceratin weighting schemes, the distance between two topics in the directory will have an impact on their relevance value.

6 Conclusions

This paper addressed the problem of inferring relevance relations between topics in a Web Directory Graph by looking at structural features of the graph only. We proposed nine different models of relevance propagation and computed them for a huge graph consisting of more than half a million nodes. This resulted in a challenging computational task, for which we implemented dedicated efficient algorithms. The resulting models were analyzed from both a quantitative and qualitative perspective.

While some models appear to better approximate the notion of relevance than others, certain general difficulties appear to rule out the possibility of defining precise models of relevance propagation by considering structural aspects only. This result has interesting practical and theoretical consequences as many existing methods attempt to identify implicit semantic relations in network representations by looking only at the structure or topology of the network (e.g., [17,19]). This calls for the investigation and development of mechanisms that integrate structural aspects with other aspects (such as content) to derive enhanced models of relevance propagation.

To the authors' knowledge, this is the first attempt to model the notion of relevance in a Web Directory Graph. However, this problem is related to the problem of estimating semantic similarity in a network representation of words, concepts or topics. Many proposals have addressed this problem by computing path distances between the nodes in the network (e.g. [19]). Other proposals estimate semantic similarity in a taxonomy based on the notion of information content [20,11]. These measures of semantic similarity have several desirable properties and a solid theoretical justification. However, because they are defined for taxonomies organized in a tree structure, they fail to capture many semantic relationships induced by the non-hierarchical components ("symbolic" and "related" links) existing in a Web Directory Graph such as ODP. More recently, an information theoretic measure of semantic similarity that accounts for both the hierarchical and non-hierarchical components of ODP has been proposed in [13] with satisfactory results. Models of relevance propagation have also been used to improve Web search. Some of these models combine content information with the link structure available in hypertext collections to rank webpages [21,8]. Other models improve the content representation of a page by applying content-based propagation between webpages through the Web structure [18].

Computational models of relevance propagation do not need to be limited to topic ontologies and Web search. Identifying relatedness relations in other ontologies requires appropriate mechanisms to model different kinds of ontology components and their

interactions. For example, the Gene Ontology[5] has two kinds of hierarchical edges ("is-a" and "part-of"). On the other hand, the WordNet ontology[6] has a much richer typology of relations. This includes semantic relations between synsets (synonym sets) such as hypernym, hyponym, meronym and holonym as well as lexical relations between senses of words (members of synsets) such as antonym, "also see", derived forms and participle.

The applicability of the proposed models of relevance propagation to the area of Artificial Intelligence and Information Science is extensive and multifarious. Since much of a reasoner's knowledge can be expressed in terms of relevance relations, a computational model of relevance propagation is a useful tool for the design of common-sense reasoning and decision-making tools .

Information Science research also builds much of its results on the notion of relevance [16]. In this field, relevance often refers to the extent to which the topic of a result matches the topic of the query. The most widely used performance measures for evaluating information retrieval methods are defined in terms of relevance. In traditional approaches users or hired evaluators provide manual assessments of relevance. However, these approaches are neither efficient nor reliable since they do not scale with the complexity and heterogeneity of available digital information. Editor-driven Web Directories have enabled the design of automatic evaluation frameworks [3,12] where the relevance of a document to a query is determined by the category of the document in the topic directory. These evaluation frameworks can highly benefit from the use of relevance propagation models.

As part of our future work we plan to improve the proposed models with additional features extracted from the websites associated with the topics and to perform controlled user studies to gain further insight into which of the models are the most accurate predictors of human judgements of relevance.

References

1. R. Akavipat, L.-S. Wu, F. Menczer, and A. G. Maguitman. Emerging semantic communities in peer web search. In *Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, P2PIR '06, pages 1–8, New York, NY, USA, 2006. ACM.
2. P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 667–674, New York, NY, USA, 2008. ACM.
3. S. M. Beitzel, E. C. Jensen, A. Chowdhury, and D. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 17–23, New York, NY, USA, 2003. ACM.
4. I. Biro, A. Benczur, J. Szabo, and A. Maguitman. A comparative analysis of latent variable models for web page classification. In *Proceedings of the 2008 Latin American Web Conference*, pages 23–28, Washington, DC, USA, 2008. IEEE Computer Society.
5. R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Inf. Process. Manage.*, 28:619–627, July 1992.

[5] `http://www.geneontology.org/`
[6] `http://wordnet.princeton.edu/`

6. S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. The structure of broad topics on the Web. In *Proceedings of the 11th international conference on World Wide Web*, WWW '02, pages 251–262, New York, NY, USA, 2002. ACM.
7. S. Chakrabarti, M. van den Berg, and B. E. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31:1623–1640, 1999.
8. I. Chibane and B.-L. Doan. Relevance propagation model for large hypertext document collections. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pages 585–595, Paris, France, France, 2007. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.
9. S. Gauch, A. Chandramouli, and S. Ranganathan. Training a hierarchical classifier using inter document relationships. *J. Am. Soc. Inf. Sci. Technol.*, 60:47–58, January 2009.
10. A. Kandel. *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley, Boston, MA, USA, 1986.
11. D. Lin. An Information-theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Franciso, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
12. A. G. Maguitman, R. L. Cecchini, C. M. Lorenzetti, and F. Menczer. Using Topic Ontologies and Semantic Similarity Data to Evaluate Topical Search. In C. von Lücken, M. E. García, and C. Cappo, editors, *XXXVI Conferencia Latinoamericana de Informática*, Asunción, Paraguay, October 2010. Centro Latinoamericano de Estudios en Informática, Facultad Politécnica – Universidad Nacional de Asunción and Universidad Autónoma de Asunción.
13. A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 107–116, New York, NY, USA, 2005. ACM.
14. B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 641–650, New York, NY, USA, 2009. ACM.
15. F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, 4(4):378–419, November 2004.
16. S. Mizzaro. Relevance: the whole history. *J. Am. Soc. Inf. Sci.*, 48:810–832, September 1997.
17. T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
18. T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen, and W.-Y. Ma. A study of relevance propagation for web search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 408–415, New York, NY, USA, 2005. ACM.
19. R. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.
20. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
21. A. Shakery and C. Zhai. A probabilistic relevance propagation model for hypertext retrieval. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 550–558, New York, NY, USA, 2006. ACM.